

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-161348

(43)Date of publication of application : 21.06.1996

(51)Int.Cl.

G06F 17/30

G06F 17/21

(21)Application number : 06-298237

(71)Applicant : CANON INC

(22)Date of filing : 01.12.1994

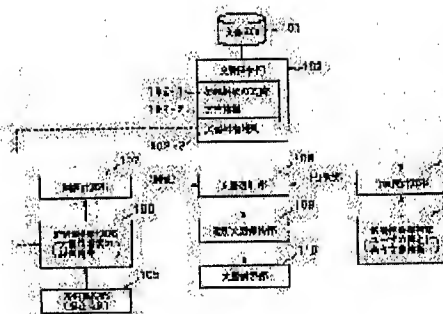
(72)Inventor : UEDA TAKANARI
HIROTA MAKOTO
ITO SHIRO
SHIBATA SHOGO
IKEDA YUJI
FUJITA MINORU

(54) DOCUMENT FILTERING METHOD AND DOCUMENT PROCESSOR

(57)Abstract:

PURPOSE: To provide a document filtering method and a document processor which filters a document by considering not only coincidence of the contents of the document and user's interest but the newness of the document.

CONSTITUTION: In the document filtering method and the document processor which filters and presents a received document, the document which should be presented is sorted by a previously set identifier and the passing time of the received document (108) and the sorted document is presented (110). At the time of sorting the representing document, the identifier of the received document and the previously set identifier are calculated (104) and based on the comparison between a threshold set corresponding to the passing time and the coincidence, whether to present the received document or not is judged. In another way, a score is given to the received document corresponding to the coincidence and the passing time and based on the comparison between the score and the prescribed threshold value, whether to present the document or not is judged. At this time, the passing time is from the preparation of the document or the reception of it, or both from preparation and from reception.



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-161348

(43) 公開日 平成8年(1996)6月21日

(51) Int.Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 17/30 17/21		9194-5L 9288-5L 9288-5L	G 0 6 F 15/ 403 15/ 20 G 0 6 F 15/ 20	3 5 0 A 5 7 0 N 5 8 6 B
審査請求 未請求 請求項の数13 O L (全 7 頁) 最終頁に続く				

(21) 出願番号 特願平6-298237

(22) 出願日 平成6年(1994)12月1日

(71) 出願人 000001007

キヤノン株式会社

東京都大田区下丸子3丁目30番2号

(72) 発明者 上田 隆也

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

(72) 発明者 廣田 誠

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

(72) 発明者 伊藤 史朗

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

(74) 代理人 弁理士 大塚 康徳 (外1名)

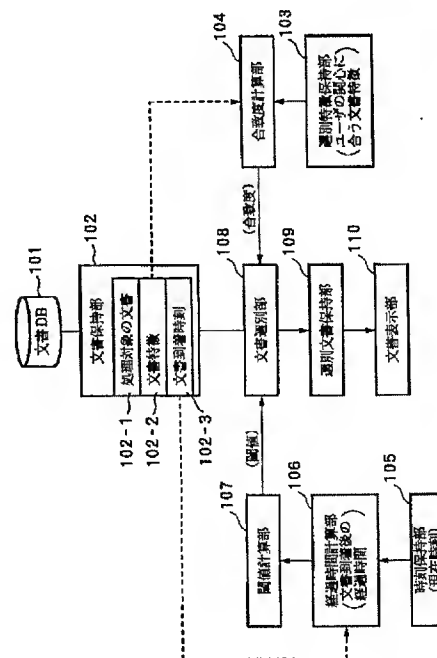
最終頁に続く

(54) 【発明の名称】 文書フィルタリング方法及び文書処理装置

(57) 【要約】

【目的】 文書の内容とユーザの関心との合致度だけでなく、文書の鮮度(新しさ)も考慮してフィルタリングを行うようにした文書フィルタリング方法及び文書処理装置を提供することを目的とする。

【構成】 受信した文書をフィルタリングして提示する文書フィルタリング方法及び文書処理装置において、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別(108)し、該選別された文書を提示(110)する。前記提示文書の選別では、受信文書の識別子と前記予め設定された識別子との合致度を計算(104)し、前記経過時間に応じて設定される閾値と前記合致度との比較に基づいて、受信文書の提示/非提示を判定する。または、前記合致度と経過時間とに応じて受信文書にスコアを与え、前記スコアと所定の閾値との比較に基づいて、受信文書の提示/非提示を判定する。ここで、前記経過時間は、前記受信文書の作成からまたは受信からのいずれか、又は作成から及び受信からの双方である。



【特許請求の範囲】

【請求項1】 受信した文書をフィルタリングして提示する文書フィルタリング方法において、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別し、該選別された文書を提示することを特徴とする文書フィルタリング方法。

【請求項2】 前記提示文書の選別は、受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、前記経過時間に応じて設定される閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する工程とを含むことを特徴とする請求項1記載の文書フィルタリング方法。

【請求項3】 前記閾値は更に文書数に応じて設定されることを特徴とする請求項2記載の文書フィルタリング方法。

【請求項4】 前記提示文書の選別は、受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、前記合致度と経過時間とに応じて受信文書にスコアを与える工程と、前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する工程とを含むことを特徴とする請求項1記載の文書フィルタリング方法。

【請求項5】 前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされることを特徴とする請求項2または4記載の文書フィルタリング方法。

【請求項6】 前記特徴ベクトルは、複数のキーワードまたは文から成ることを特徴とする請求項5記載の文書フィルタリング方法。

【請求項7】 前記経過時間は、前記受信文書の作成からまたは受信からのいずれか、又は作成から及び受信からの双方であることを特徴とする請求項1または2または4記載の文書フィルタリング方法。

【請求項8】 受信した文書をフィルタリングして提示する文書処理装置において、受信文書を作成からの経過時間及び／又は受信からの経過時間と共に保持する文書保持手段と、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別する選別手段と、該選別された文書を提示する提示手段とを備えることを特徴とする文書処理装置。

【請求項9】 前記選別手段は、受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、前記経過時間に対応する閾値を設定する閾値設定手段と、前記閾値と前記合致度との比較に基づいて、受信文書の

提示／非提示を判定する判定手段とを含むことを特徴とする請求項8記載の文書処理装置。

【請求項10】 前記閾値設定手段は、更に文書数にも応じて閾値を設定することを特徴とする請求項9記載の文書処理装置。

【請求項11】 前記選別手段は、受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、前記合致度と経過時間とに応じて受信文書にスコアを与えるスコア計算手段と、

前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含むことを特徴とする請求項8記載の文書処理装置。

【請求項12】 前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされることを特徴とする請求項9または11記載の文書処理装置。

【請求項13】 前記特徴ベクトルは、複数のキーワードまたは文から成ることを特徴とする請求項12記載の文書処理装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は文書処理装置に関し、特にユーザのもとに入ってくる文書のうちユーザが関心を持つ文書を選別し、その結果を出力する文書フィルタリング方法及び文書処理装置に関するものである。

【0002】

【従来の技術】昨今、記録媒体の大容量化と低価格化、また、ワードプロセッサの普及などによって、電子化された文書の量が增大している。さらに、ネットワークの整備が進み、電子メールや電子ニュースなどのように、メディアを介して直接ユーザのもとに届く電子化文書の量も増えている。このため、ユーザが処理できる量を越えた文書が入ってくるようになり、本当に必要な情報が不要な情報の中にも埋もれてしまうという、いわゆる「情報洪水」が問題になってきている。

【0003】この問題への対応策として、ユーザが関心を持つであろう文書のキーワードを自動的に選別する「文書フィルタリング」の技術が用いられるようになってきた。「文書フィルタリング」では、予め文書に対するユーザの関心をキーワードや文書の内容として設定しておき、この設定と送られてきた文書の内容を比較して、一致がある場合にはその文書をユーザに見せ、一致していないときにはその文書をユーザに見せないという制御をしている。すなわち、「文書フィルタリング」の技術によって、ユーザは自分にとって関心のないような文書を最初から見なくてすむようになり、時間を有効に利用できるようになった。

【0004】

【発明が解決しようとする課題】しかしながら、フィル

タリングされた文書を定期的にユーザが見ている間は問題ないが、ユーザがしばらく文書を見ることができずにフィルタリングされた文書が溜まってしまったような場合、文書数が多くなり、結局重要な情報を見落としてしまう可能性が出てくる。

【0005】従って、重要な情報を見落とさないための対策として、フィルタリングされた文書をユーザの関心のあるキーワードや文書の内容との合致の度合（合致度）の順に整列して表示し、合致度の高いものから見るようにすることも行われるが、この場合、最新の必要文書が後に来る等、文書の「鮮度」が考慮されていないことに問題がある。

【0006】電子メールのように日々ユーザのもとに届くような文書については、鮮度が重要で、一般に「新しい文書ほど情報の価値がある」ということが言える。届いたときには重要な情報であっても、時間の経過と共に重要度が下がる文書も多い。しかし、文書フィルタリングにおいて単に文書の内容だけを考慮したのでは、こうしたことが反映されない。

【0007】従来装置には、以上のような種々の問題があった。本発明は、上述した従来の課題を解決し、文書の内容とユーザの関心との合致度だけでなく、文書の鮮度（新しさ）も考慮してフィルタリングを行うようにした文書フィルタリング方法及び文書処理装置を提供することを目的とする。

【0008】

【課題を解決する為の手段】上述の課題を解決するために、本発明の文書フィルタリング方法は、受信した文書をフィルタリングして提示する文書フィルタリング方法において、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別し、該選別された文書を提示することを特徴とする。

【0009】ここで、前記提示文書を選別は、受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、前記経過時間に応じて設定される閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する工程とを含む。また、前記閾値は更に文書数に応じて設定される。また、前記提示文書を選別は、受信文書の識別子と前記予め設定された識別子との合致度を計算する工程と、前記合致度と経過時間とに応じて受信文書にスコアを与える工程と、前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する工程とを含む。また、前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされる。また、前記特徴ベクトルは、複数のキーワードまたは文から成る。また、前記経過時間は、前記受信文書の作成からまたは受信からのいずれか、又は作成から及び受信からの双方である。

【0010】又、本発明の文書処理装置は、受信した文

書をフィルタリングして提示する文書処理装置において、受信文書を作成からの経過時間及び／又は受信からの経過時間と共に保持する文書保持手段と、予め設定された識別子と受信文書の経過時間とから提示すべき文書を選別する選別手段と、該選別された文書を提示する提示手段とを備えることを特徴とする。

【0011】ここで、前記選別手段は、受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、前記経過時間に対応する閾値を設定する閾値設定手段と、前記閾値と前記合致度との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含む。また、前記閾値設定手段は、更に文書数にも応じて閾値を設定する。また、前記選別手段は、受信文書の識別子と前記予め設定された識別子との合致度を計算する合致度計算手段と、前記合致度と経過時間とに応じて受信文書にスコアを与えるスコア計算手段と、前記スコアと所定の閾値との比較に基づいて、受信文書の提示／非提示を判定する判定手段とを含む。また、前記識別子は特徴ベクトルで表わされ、前記合致度は受信文書の特徴ベクトルと前記予め設定された特徴ベクトルとの内積で表わされる。また、前記特徴ベクトルは、複数のキーワードまたは文から成る。

【0012】

【作用】以上の構成により、文書フィルタリングの際に、文書特徴と選別特徴との合致度が、経過時間に応じて定まる閾値を越えた場合に、その文書を必要と判断してユーザに呈示することにより、文書の鮮度を考慮したフィルタリングを行うことが可能となる。

【0013】

【実施例】以下、本発明の実施例を添付図面を用いて詳細に説明する。図1は、本実施例の文書処理装置の処理の論理構成を示すブロック図である。図1において、101は、ユーザの元に入ってきた文書を格納している文書データベースである。102は、処理対象の文書102-1と、その文書特徴102-2と、文書到着時刻102-3を保持する文書保持部である。103は、ユーザの関心に合う文書の文書特徴（選別特徴）を保持する選別特徴保持部である。104は、処理対象の文書の文書特徴と選別特徴との合致度を計算する合致度計算部である。105は、現在の時刻を保持する時刻保持部である。106は、文書が到着してからの経過時間を計算する経過時間計算部である。107は、経過時間をもとに閾値を計算する閾値計算部である。108は、合致度と閾値の関係によって文書を選択する文書選択部である。109は、文書選択部108で選ばれた文書を保持する選別文書保持部である。110は、選別文書保持部109に保持された文書を表示する文書表示部である。

【0014】図2は本実施例の文書処理装置のハードウェア構成を示す図である。図2において、201は、図1及び図4に示す制御手順300を記憶する制御メモリ

である。これはROMであってもよいし、RAMであっても良い。202は、制御メモリ201に記憶されている制御手段に従って処理を行う中央処理装置である。203はメモリで、上記文書保持部102、選別特徴保持部103、選別文書保持部109を有する。なお、時刻保持部105は現在の時刻を保持するところで、ハードウェアであっても良い。204はキーボード、208はポインティングデバイスであり、操作者が操作する。205はディスクであり、文書データベース101を有する。206はディスプレイで、CRTであってもよいし、液晶ディスプレイであってもよい。これは文書を表示するのに用いる。207は各構成要素を接続する為のバスである。209はフロッピーである。

【0015】図3は、図2に示す制御メモリ201の中にある制御手順300の構成を更に示すものであり、合致度計算部104と、経過時間計算部106と、閾値計算部107と、文書選別部108と、文書表示部110とを含む。図4は図3に示した処理部に対応する動作手順を示すフローチャートである。図4を参照しながら、本発明の一実施例の動作を説明する。なお、本実施例では文書の特徴の表現方法として、一般に知られているベクトル空間モデルを利用する。

【0016】ベクトル空間モデルでは、文書の特徴を表現するのに、N個のキーワードを用意し、文書毎に各キーワードの重みを設定する。これは、N次元空間のベクトルとみなすことができる。このベクトルを長さ1に正規化する。文書の特徴の合致度はそれぞれのベクトルの内積として表す。まず、ステップS301で、合致度計算部104において、文書保持部102に保持された文書特徴102-2と、選別特徴保持部103に保持された選別特徴との合致度を計算する。先に述べたように、合致度は文書特徴を表すベクトル間の内積で表すので、文書特徴をベクトルd、選別特徴をベクトルsとすると、合致度は $(d \cdot s)$ になる。

【0017】ステップS302では、経過時間計算部106において、文書が到着してから経過時間を計算する。これは、時刻保持部105に保持された現在時刻と、文書保持部102に保持された文書到着時刻102-3との差分によって求めることができる。ステップS303では、閾値計算部107において、ステップS302で計算した経過時間から閾値を計算する。閾値は経過時間tに従って決めるが、tの増加と共に増加するような関数 $f(t)$ であればどのような決めかたをしても構わない。例えば、tを日数(端数切り捨て)で表現し、 $f(t) = 1 - 1/(t+2)$ のようにする。

【0018】ステップS304では、ステップS301で計算した合致度と、ステップS303で計算した閾値を比較する。合致度が閾値を越えていない場合はそのまま処理を終了する。ステップS304で合致度が閾値を越えていればステップS305に進み、文書選別部10

8で、この文書を選択して選別文書保持部109に保持する。そして処理を終了する。

【0019】例えば、ある文書の文書特徴ベクトルと選別特徴ベクトルの内積pを0.78とする。この文書が到着してから2日経過した時点では $f(t) = 1 - 1/(t+2)$ に代入すると、 $f(2) = 0.75$ であり、 $p > f(2)$ となって、文書を選択するが、3日経過した時点では $f(3) = 0.8$ であり、 $p < f(3)$ となって、文書を選択しないことになる。

【0020】次に、文書表示部110により、選択された文書を表示する。尚、前記実施例では、文書をユーザに呈示する際に合致度と閾値の計算を両方行っているが、合致度の計算を文書が届いた時点で行ない、その値を保存しておくようにしておき、次に、文書をユーザに呈示する際には閾値の計算だけを行うようにする。こうすることによって、文書をユーザに呈示する際の処理時間を短縮することができる。

【0021】又、前記実施例では、文書の特徴にベクトル空間モデルを利用したが、この表現方法に限らない。すなわち、文書特徴と選別特徴の合致度を計算し、それに対する閾値を経過時間tの関数として定義できれば良い。又、前記実施例では閾値を計算する際に経過時間を日数で表わしたが、秒・分・時間などどんな単位で表現しても構わない。また閾値の関数も上記実施例に挙げたものに限るものではない。

【0022】又、前記実施例では、経過時間tの関数で定まる閾値を用意して合致度との大小を調べたが、文書の選別特徴の合致度と経過時間tから定まるスコアの合計を文書選別のスコアとし、文書選別のスコア自体に文書特徴の合致度だけでなく、経過時間tも反映するようにし、更に、経過時間tの増加でスコアが減少するように定めれば良い。このスコアに対して一定の閾値を設けて文書を選別するものでもよい。

【0023】又、前記実施例では、溜まった文書数を考慮していない。しかし、閾値を計算する際に文書数も反映して、文書数が多い場合にはそれだけ閾値を高くするようにしても良い。こうすることによって、文書数が多い場合でも一定の数の文書をユーザに提示することができる。又、前記実施例では、文書が到着してから経過時間を用いたが、文書が作成された時刻がわかる場合は、作成されてからの経過時間を用いるようにしても良いし、作成と受信とを共に用いてもよい。

【0024】更に、本発明は、複数の機器から構成されるシステムに適用しても、1つの機器から成る装置に適用しても良い。また、本発明はシステム或は装置にプログラムを供給することによって達成される場合にも適用できることはいうまでもない。

【0025】

【発明の効果】以上説明したように、本発明によれば、より新しい情報をより重要なものとして文書フィルタリ

ングを行うようにしたので、限られた時間でユーザが情報を効率的に利用することができるという効果がある。

【図面の簡単な説明】

【図 1】本実施例の文書処理装置の処理の論理構成を示すブロック図である。

【図 2】本実施例の文書処理装置のハードウェア構成を示す図である。

【図 3】本実施例の制御手順の構成を示す図である。

【図 4】本実施例に係る動作手順を示すフローチャートである。

【符号の説明】

101 文書データベース

102 文書保持部

102-1 処理対象の文書

102-2 文書の特徴

102-3 文書到着時刻

103 選別特徴保持部

* 104 合致度計算部

105 時刻保持部

106 経過時間計算部

107 閾値計算部

108 文書選別部

109 選別文書保持部

110 文書表示部

201 制御メモリ

202 中央処理装置

10 203 メモリ

204 キーボード

205 ディスク

206 ディスプレイ

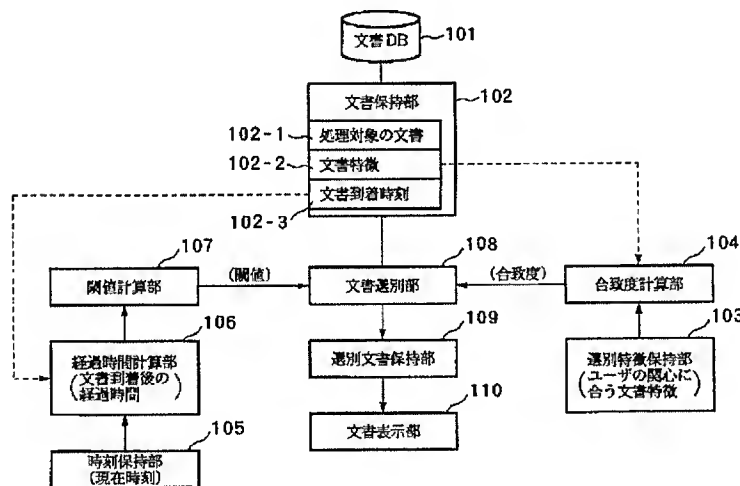
207 バス

208 ポインティングデバイス

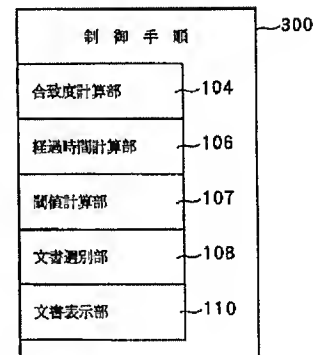
209 フロッピー

*

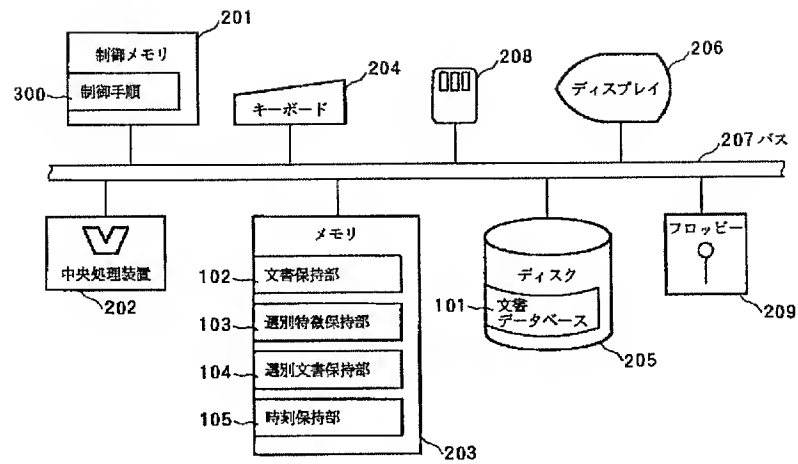
【図 1】



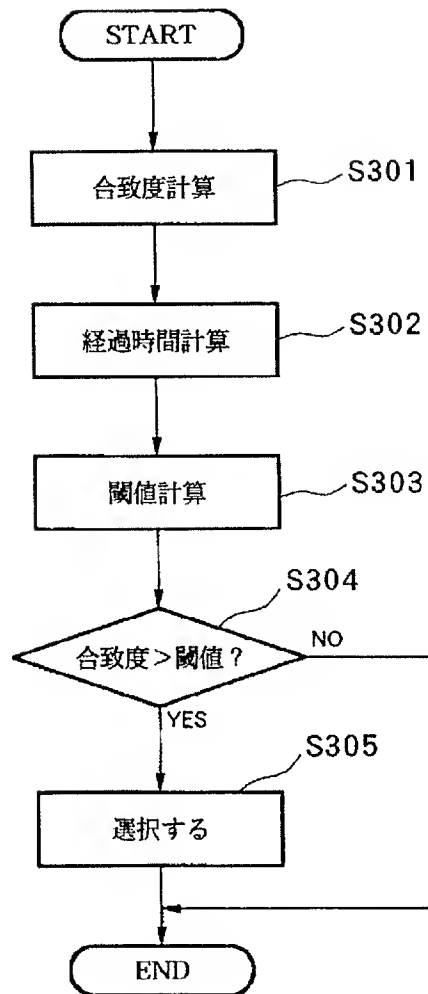
【図 3】



【図 2】



【図4】



フロントページの続き(51)Int.Cl.⁶

識別記号

片内整理番号

F I

技術表示箇所

9194-5L

15/40

310 F

(72)発明者 柴田 昇吾

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

(72)発明者 池田 裕治

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内

(72)発明者 藤田 稔

東京都大田区下丸子3丁目30番2号 キヤ
ノン株式会社内